Chatbots in Data Science

Using chatbots for different parts of a data science exercise

May 2025

By Dwayne Phillips

Summary

This essay contains a description of an experiment in data science that uses chatbots as tools for several parts of the experiment. This is not a polished experiment, but rather a first step in exploring the use of chatbots in data science. Chatbots are useful. As different vendors develop their chatbot systems, further experiments will demonstrate their use in general data science.

Introduction

This essay documents an experiment in data science using chatbots. The purpose is to learn if chatbots can assist in finding specific information that is contained in large, unstructured text files. This can all be done manually, but the time required is prohibitively large.

The sample task is to download a body of emails and query them to create a useful body of knowledge about the people mentioned in the emails.

This experiment uses Google Gmail, Google Takeout, Google Gemini, OpenAI ChatGPT, Python, and Google NotebookLM. This is not an endorsement of any of these tools. They were used for convenience

The actual content of the email will not be revealed to protect the privacy of persons.

A further experiment would be to download widely published text such as newspaper archives as a subject. That would allow publication of details and allow for a better judgement of chatbot performance (% of information found, etc.).

Downloading Email

The emails used in this experiment were from the author's Gmail account. Google has a tool at Takeout.google.com that will download all emails on record. For this experiment,

only the emails that were filed in one folder were downloaded. Takeout downloaded a 900MB file in the mbox format (see https://en.wikipedia.org/wiki/Mbox).

Reducing the Downloaded Email and Changing the File Format

The downloaded email file was too large to work with for this experiment. It also contained information that was not of interest. The narrow the scope of the emails, a Python program would read through the mbox file, grab only emails with a given text in the Subject line, and place those into a text file.

A chatbot was asked to create the Python program. The only editing of the Python program was to enter the filename of the mbox file. The Python program ran as needed and created a text file containing the emails with the specified text in the Subject line.

Creating a PDF File

The various chatbots available all read PDF files an inputs. A simple way to convert the text file to PDF is to read it into Microsoft Word and have word “print” the file to a PDF file.

Using Chatbots to Gather Information

The email contents in the PDF file contain several MB of unstructured information about people. Attaching this PDF to different chatbots allow for prompting to collect information about each individual mentioned. This information finding could be performed manually, but that is a tedious, error-prone task.

Three different chatbots were tried to pull information from the email-containing PDF. This essay is not a test of chatbot performance comparing one to another. It is a demonstration that today’s chatbots can pull information from large text files.

Today’s chatbots can pull information about person’s named in prompts. Another prompting experiment was to list the names of a dozen persons in a second file, ask the chatbot to read each name in the second file, and produce information about each person. The chatbots were able to do that.

Conclusions

Chatbots are new and useful tools in data science. They can create tools such as Python programs as well as retrieve useful information from large text files.

There is much further work to be done as experiments.